RECENT ADVANCES IN HNC'S CONTEXT VECTOR INFORMATION RETRIEVAL TECHNOLOGY

Marc R. Ilgen, David A. Rushall
HNC Software, Inc., 5930 Cornerstone Court West, San Diego, CA 92121 USA
email: mri@hnc.com, dar@hnc.com

ABSTRACT

Over the past few years, HNC has developed a neural network based, vector space approach to text retrieval. This approach, embodied in a system called MatchPlus, allows the user to retrieve information on the basis of meaning and context of a free text query. The MatchPlus system uses a neural network based, constrained self-organization technique to learn word stem interrelationships directly from a training corpus, thereby eliminating the need for hand crafted linguistic knowledge bases and their often substantial maintenance requirements. This paper presents results from recent enhancements to the basic MatchPlus concept. These enhancements include the development of a one step learning law that greatly reduces the amount of time and/or computational resources required to train the system, and the development of a prototype multilingual (English and Spanish) text retrieval system.

1. INTRODUCTION

While the current MatchPlus learning law has proven to be effective in encoding relationships between words, it is computationally intensive and requires multiple passes through the training corpus. The purpose of the one step learning law is to approximate the behavior of the original learning law while performing only a single pass through the training corpus. The one step learning law uses a single pass through the training corpus to obtain desired dot product values for the set of trained context vectors. The desired dot product values are determined on the basis of information theoretic statistical relationships between co-occurring word stems found in the training corpus. Desired dot products are found such that words that tend to co-occur will have context vectors that point in similar directions while words that do not co-occur will have context vectors that tend to be orthogonal. These desired dot products are used to perform a quasi-linear transformation on an initial set of quasi-orthogonal, high dimensional vectors. This vector transformation and subsequent renormalization results in a set of context vectors that represents the relationships between word stems in a near-optimal fashion. The time requirements for training this set of vectors scale as O(Nn) where N is the number of word stems in the vocabulary and n is the average number of word stems found to co-occur (and/or be related to) any given word stem (usually on the order of several hundred). This new learning law reduces the training time by a factor on the order of 100 over the original context vector learning law with little or no degradation in performance. Results can be improved even further by adjusting the dimension of the context vectors.

HNC has also developed an approach to learning stem-level relationships across multiple languages and has used this approach to develop a prototype multilingual retrieval system. This technique, called "symmetric learning", is based upon the use of tie words, which provide connectivity between each language's portion of the context vector space. In the symmetric approach, learning is conducted using both languages simultaneously, thus removing any donor language biases. Tie words are used to connect the context vector space for multiple languages through a "unified hash table". The unified hash table provides the mechanism to translate a stem into an associated context vector. In the English-only MatchPlus system, this is a straight forward process. The stem in question is fed to the hashing function and the index is produced. The resulting index is the offset in the stem hash table. The content of that location in the hash table is a pointer to the context vector data structure. Using this approach (hash function collisions not withstanding), each unique stem results in a unique entry and thus a unique context vector. In the multilingual system, a tie word list is used to provide multiple references, one word stem from each language, for common context vectors. Context vector learning is performed in multiple languages

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **MAY 1996**	2. REPORT TYPE	3. DATES COVERED **00-00-1996 to 00-00-1996**

4. TITLE AND SUBTITLE **Recent Advances in HNC’s Context Vector Information Retrieval Technology**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **HNC Software, Inc,5930 Cornerstone Court West,San Diego,CA,92121**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996. Sponsored by the Defense Advanced Research Projects Agency.

14. ABSTRACT

Over the past few years, HNC has developed a neural network based, vector space approach to text retrieval. This approach, embodied in a system called MatchPlus, allows the user to retrieve information on the basis of meaning and context of a fi’ee text query. The MatchPlus system uses a neural network based constrained sel~organization technique to learn word stem interrelationships directly 3~om a training corpus, thereby eliminating the need for hand crafted linguistic knowledge bases and their often substantial maintenance requirements. This paper presents results fi’om recent enhancements to the basic MatchPlus concept. These enhancements include the development of a one step learning law that greatly reduces the amount of time and~or computational resources required to train the system, and the development of a prototype multilingual (English and Spanish) text retrieval system.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**10**	

simultaneously using multilingual training corpora. HNC has performed preliminary evaluation of an English-Spanish version of this system by examining stem trees for tie words and non-tie words. Results indicate that the English-Spanish MatchPlus prototype is able to learn reasonable word stem interrelationships for tie words and non-tie words, thereby demonstrating the suitability of this concept for further development.

2. TECHNICAL BACKGROUND

The HNC MatchPlus system was developed as part of the ARPA-sponsored TIPSTER text program. MatchPlus uses an information representation scheme called *context vectors* to encode similarity of usage. Other vector space approaches to text retrieval exist, but none embody the ability to learn word-level relationships [1-5]. Key attributes of the context vector approach are as follows: During this effort, the initially proposed context vector approach using human defined coordinates and initial conditions was extended and refined to allow fully automatic generation of context vectors for text symbols (stems) based upon their demonstrated context of usage in training text. The MatchPlus system learns relationships at the stem level and then uses those relationships to construct a context vector representation for sets of symbols. For the text case, these sets of symbols are paragraphs, documents and queries.

To start the learning process, each stem is associated with a random vector in the context vector space. Random unit vectors in high dimensional floating point spaces have a property that is referred to a "quasi-orthogonality"[6]. That is, the expected value of the dot product between any pair of random context vectors selected from the set is approximately equal to zero (i.e. all vectors are approximately perpendicular to one another). This property of quasi-orthogonality is important because it serves as the initial condition for the context vector learning algorithm. The usage of the context vector technique is predicated upon the rule that symbols (stems) that are used in a similar context (exhibit proximate co-occurrence behavior) will have trained vectors that point in similar directions. Conversely, stems that never appear in a similar context will have context vectors that are approximately orthogonal.

To achieve the desired representation, the context vector learning algorithm must take the context vectors for symbols that co-occur and move them toward each other. Symbols that do not co-occur are left in their quasi-orthogonal original condition. It is a basic tenet of the MatchPlus approach that "words that are used in a similar context convey similar meaning". Since the learning is driven by proximate co-occurrence of words, the learning results in a vector set *where closeness in the space is equivalent to closeness in subject content.* To perform learning, a learning window is used to identify local context. The window is "slid" through each document in the corpus. The window has 1 target stem and multiple neighbor stems. Once the context window has been determined, the learning rule of "Move context vector for target in the direction of the context vector of the neighbors" is applied. Once the correction is made, we move the learning window to next location and the learning operation is repeated. The equation for this learning is shown in Figure 1.

$$T_i^{New} = T_i^{Old} + \gamma \cdot \sum_j \left(\alpha_{ij} - T_i^{Old} \bullet N_{ij} \right) \cdot N_{ij}$$

$$T_i^{New} = \frac{T_i^{New}}{\left\| T_i^{New} \right\|}$$

where:

T_i^{New} = Context vector of target *i* after update

T_i^{Old} = Context vector of target *i* before update

γ = Adjustment step size

N_{ij} = Context vector for neighbor *j* of target *i*

α_{ij} = Desired context vector dot product for target *i* and neighbor *j*

Figure 1. MatchPlus Learning Equations

Several points should be noted:

- All stem vectors are of length 1 (unit vectors). In this paradigm, only the direction of the vector carries information.
- Fully trained vectors have the property that words that are used in a similar context will have vectors that point in similar directions as measured by the dot product.
- Words that are never used in a similar context will retain their initial condition of quasi-orthogonality. That is, approximately orthogonal with a dot product of approximately zero.
- Trained context vectors result in a concept space where similarity of direction corresponds to similarity of meaning.
- No human knowledge is required for training to occur. Only free text examples are needed.

- The algorithm determines the coordinate space of the context vectors.

When the training is complete, "words that are used in a similar context will have their associated vectors point in similar directions". Conversely, words that are never used in a similar context will have vectors that are approximately orthogonal.

At the summary level, the MatchPlus system translates free text into a mathematical representation in a meaningful way. Note that the MatchPlus approach does not use any external dictionaries, thesauri or knowledge bases to determine word vector relationships. These relationships are learned automatically using only the text examples provided for learning. The result of the learning procedure is a vocabulary of stem context vectors that can be used for a variety of applications including document retrieval [7], routing [8], document clustering and other text processing tasks.

Once the stem learning is complete, it is possible to "query" the vector set to determine the nature of the learned relationships. To perform this operation, the user selects a "root" word and the trained context vector for that word is determined by a table lookup in the context vector vocabulary. MatchPlus computes the dot product of every other word vector in the vocabulary to the selected word. The resulting dot products are sorted by magnitude where larger means closer in usage.

Sets of words (text passages and queries) and documents can also be represented by context vectors in the same information space. Document context vectors are derived as the inverse document frequency-weighted sum of the context vectors associated with words in the document. Document context vectors are normalized to prevent long documents from being favored over short documents. The resulting document context vectors have the property that *documents that discuss similar themes will have context vectors that point in similar directions*. It is this property that translates the problem of assessment of similarity of content for text into a geometry problem. Documents that are similar are close in the space and dissimilar documents are far away. Additionally, it should be noted that all document vectors are unit length. This prevents system biases in retrieval due to document length.

3. ONE STEP CONTEXT VECTOR LEARNING

The sections below describe an approach to context vector learning that greatly reduces the amount of computer time and resources required to obtain a trained set of stem context vectors. This approach uses a single pass through the training corpus (or corpora) to obtain desired dot product values for the set of trained context vectors. These desired dot products are used in a single pass through the vocabulary of word stems to expand a starting set of quasi-orthogonal, high dimensional vectors. This vector expansion and subsequent renormalization results in a set of context vectors that represents the relationships between words stems in a near-optimal fashion. The time requirements for training this set of vectors scale as O(Nn) where N is the number of word stems in the vocabulary and n is the average number of word stems found to co-occur (and/or be related to) any given word stem (usually on the order of several hundred). Using a near-worst case estimate of n=1000 word stems, a vocabulary size of 50,000 words, and assuming that at least ten iterations of the original learning law are required for convergence (more often at least one hundred iterations are required), this new learning law reduces the training time by a factor of between 10 and 500 (depending on whether or not the non co-occurring terms are explicitly considered in the current learning law).

3.1 Current MatchPlus Context Vector Learning Law

The current MatchPlus context vector learning law is presented in Figure 1 and discussed in Section 2. This learning law can be derived as a stochastic gradient descent procedure for minimizing the cost function

$$J(T_i, T_j) = \frac{1}{2}\sum_{i,j}\left(\alpha_{ij} - T_i \bullet T_j\right)^2$$

where:

T_i = Context vector for word stem i

T_j = Context vector for word stem j

α_{ij} = Desired dot product for word stem i and j context vectors

Figure 2. Learning Law Cost Function

subject to the constraints

$$\|T_i\| = \|T_j\| = 1$$

where:

$$\|T\| = (T \bullet T)^{1/2}$$

Figure 3. Learning Law Vector Magnitude Constraints

The factors $\alpha_{i,j}$ are the desired dot products for the trained set of context vectors. These desired dot products are found as a function of co-occurrence statistics for word stems i and j. In most cases the number of words for which $\alpha_{i,j}$ is non-zero (i.e. the co-occurring words) is several orders of magnitude smaller than the size of the vocabulary. In theory, the summation on the right hand side extends over all word stems in the vocabulary. In practice, however, this summation is performed only over words that co-occur with the target word stem i. Since n=number of co-occurring words is usually much less than N=number of vocabulary word stems, summing only over co-occurring words represents a considerable time savings. Non co-occurring word stem context vectors are adjusted by subtracting the mean context vector at the end of each update iteration. This has the effect of spreading out the context vectors, hopefully driving the context vectors of non co-occurring words closer to orthogonality. With this approximation, the time requirements for the current learning law scale as O(kNn) where k is the number of iterations required for convergence.

3.2 Approach to One Step Learning

The objective of any learning law used to train context vectors is to minimize the cost function specified in Figure 2 subject to the constraints in Figure 3. In order to avoid the requirement for multiple iterations, HNC proposes to evaluate the performance of the following one step learning law:

$$T_i^{New} = T_i + \eta \sum_j \alpha_{ij} T_j$$

$$T_i^{New} = \frac{T_i^{New}}{\|T_i^{New}\|}$$

where:

T_i^{New} = Context vector of target i after one step update

T_i = Context vector of target i before update

T_j = Context vector for word stem j before update. Word stem j co-occurs with word stem i

η = Design parameter chosen to optimize performance

α_{ij} = Desired context vector dot product for target i and co-occurring stem j

Figure 4. One Step Learning Law Equations.

Note that the summation in Figure 4 is over co-occurring word stems. This learning law is motivated by the following observation. Suppose there exists a cost function of two variables x_1 and x_2, where $J(x_1,x_2) = \frac{1}{2} \sum_{i,j=1,2} (\alpha_{ij} - x_i x_j)^2$. Suppose further that we wish to choose δx_1 and δx_2 such that replacing x_1 and x_2 with the quantities $x_1 + \delta x_1$ and $x_2 + \delta x_2$ minimizes the cost function. For the situation in which $\|x_1\| = \|x_2\| = 1$ and δx_1 and δx_2 are assumed to be small, it is easily demonstrated that the solutions for δx_1 and δx_2 are $\delta x_1 = \frac{\alpha_{12}}{2} x_2$ and $\delta x_2 = \frac{\alpha_{12}}{2} x_1$. Adding these solutions to x_1 and x_2 yields an expression similar to that of Figure 4. Of course, the fact that the resulting vectors must be normalized makes the analogy only approximate. However, Figure 4 can be viewed as a one step approximation to the optimal solution. The value of this approximate solution is that it provides adequate performance with only a fraction of the computational requirements. This one step learning law scales as O(Nn), so that it is faster than the current learning law by a factor of k (number of original learning law iterations) and is faster than the theoretically derived learning law by a factor of kN/n. For reasonable values of k, N, and n, this translated into a time savings of a factor of 10 to 1000.

3.3 Summary of One Step Learning

The successful development and testing of the one step learning law offers the possibility of much faster context vector training. The performance of the system using this law can be optimized through parameter sweeps on context vector dimension and free parameter η (see Figure 4).

4. APPROACH TO MULTILINGUAL INFORMATION RETRIEVAL (MIR)

The objective of solving the MIR problem is to provide the analyst/user with a flexible high performance tool to allow retrieval of relevant information from multilingual corpora without the need for prior translation of large volumes of text.

The key issue is prior translation of the foreign language material. Clearly, if all material was translated to a uniform representation, say English, the problem is solved. However, translation is time consuming, costly and subjective. Additionally, the current volumes of information would overwhelm any organization who attempted to perform bulk translation. Machine translation efforts have been partially successful, but these techniques frequently ignore subtleties in the translation process. Additionally, the cost of development, tuning and validation of this approach is a hindrance to widespread use.

HNC has developed an approach to the MIR problem that leverages the context vector technology. It is called symmetric learning and its attributes, as well as the implications of its attributes, are discussed in the section below. Explanations of the approach will be given from the frame of reference of two simultaneous languages. However, it should be noted that these approaches are extensible to many languages being processed simultaneously. These discussions assume that language 1 is English and language 2 is Spanish. It should also be noted that HNC has implemented a minimal subset of the symmetric approach as a proof of concept. The preliminary results are extremely encouraging. A description of this system, the training corpus and the preliminary results are provided in Section 5.

4.1 Symmetric Learning

HNC has developed an approach to learning stem-level relationships across multiple languages. This technique, called "symmetric learning", is based upon the use of tie words. These tie words provide connectivity between each language's portion of the context vector space. However, learning is conducted using both languages simultaneously, thus removing any donor language biases.

The symmetric approach is based upon the use of a "unified hash table". The unified hash table provides the mechanism to translate a stem into an associated context vector. In the English-only MatchPlus system, this is a straight forward process. The stem in question is fed to the hashing function and an index is produced. The resulting index is the offset in the stem hash table. The contents of that location in the hash table is a pointer to the context vector data structure. Using this approach (hash function collisions not withstanding), each unique stem results in a unique entry and thus a unique context vector. What is proposed is to use the tie word list to provide references for common context vectors. An example is shown in Figure 5. Assume that "attack" and "ataque" have been chosen as a tie word pair. Since these words should have the same context vector, some form of connection must be made between the words. Figure 5 shows 4 words in the unified hash table: "rebel", "attack", "ataque", and "contra". Without hash table unification based upon the tie word list, all four words would have unique and independent context vectors. However, as can be seen in the figure, the hash table entries for the tie words have been forced to point to a common context vector entry. This very simple approach allows multiple references to the same context vector entity.

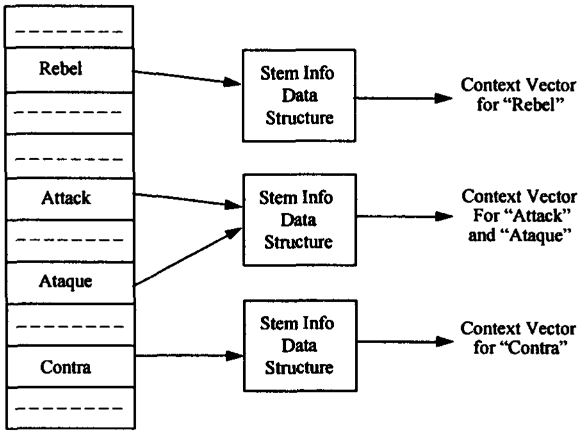

Figure 5. Unified Hash Table Example.

Once the mechanism for multiple references has been established, the next step is to consider the actual training algorithm. Example training text for English and Spanish is shown in Figure 6. For this example, it is assumed that the pair "attack" and "ataque" are a tie word pair. Note that in this example, the text chosen is a near-literal translation. There is no requirement for parallel text for the symmetric learning algorithm. The English text in Figure 6 comes from the passage, "Four people were killed in the attack by the rebel group Shining Path". The corresponding Spanish text is "Quatro personas fueron matadas en el ataque por el group contras Sendero Luminoso". Figure 6 shows the context window for the stemmed text centered on the tie word attack.

Like the standard MatchPlus context vector learning algorithm, the symmetric learning approach will utilize a convolutional "context window" with a center and neighbors. The stem at the center of the window is called the "target". The context vector for the target stem is adjusted in the direction of its neighbors' context vectors.

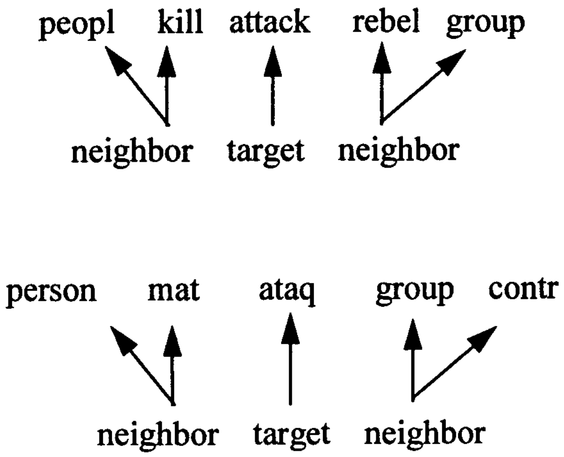

Figure 6. Symmetric Learning Example.

The steps that will occur during learning given the text example shown in Figure 6 are as follows:

- The convolutional window location is chosen and the target and neighbor stems are identified. In the English portion of this example, the window is centered on the word "attack". The neighbor words are "people", "killed", "rebel", and "group".
- The context vector for "attack" is moved in the direction of its neighbors. When the update is completed, the window is moved and the process is repeated.
- Spanish text is processed using the same approach. In the Spanish portion of this example, the window has as its center the word "ataque". Neighbors for "ataque" are "personas", "matado", "groupo", and "contra". The context vector for "ataque" is moved in the direction of its neighbors. When the update is completed, the window is moved and the process is repeated.
- Note that "attack" and "ataque" are a tie word pair. As a consequence, they share a common context vector. As a consequence, the context vector for this pair has been influenced by the words that have occurred in a similar context in *both languages*. Specifically, the attack-ataque tie word pair has been influenced by "people", "kill", "rebel", "group", "personas", "matado", "groupo", and "contra".
- Since all context vectors are in the same information space, the symmetric learning technique will result in a unified information space for both languages. Because of the "second order" learning effects of the context vector approach, not only will "attack" be related to "people" and "personas", but "people" will be related to "personas", "matado", "groupo", etc.

The block diagram for generation of a system using the symmetric approach is shown in Figure 7. As can be seen in this figure, the symmetric system build uses the unified hash table as the basis for combining the stem sets from both languages. Once this process has taken place, all stem context vectors are stored in a single dataset. This unified set of context vectors is the basis for formation of document context vectors. When the system generation is complete, MIR is ready for query processing. The block diagram for this process is shown in Figure 8.

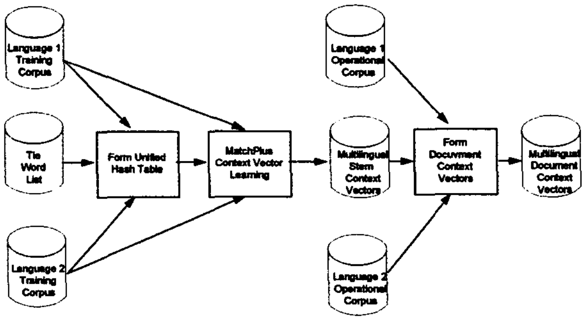
Language 1 Training Corpus
Tie Word List
Language 2 Training Corpus
Form Unified Hash Table
MatchPlus Context Vector Learning
Language 1 Operational Corpus
Multilingual Stem Context Vectors
Language 2 Operational Corpus
Form Docuvment Context Vectors
Multilingual Document Context Vectors

Figure 7. Symmetric Approach System Generation.

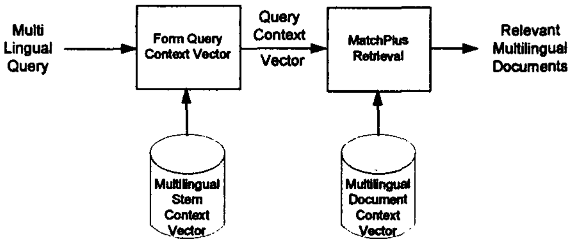
Multi Lingual Query
Form Query Context Vector
Query Context Vector
MatchPlus Retrieval
Relevant Multilingual Documents
Multilingual Stem Context Vector
Multilingual Document Context Vector

Figure 8. Symmetric Approach Query Processing.

Attributes of the symmetric learning approach are as follows:

1. Once tie word pairs (or n-tuples) have been selected, all subsequent processing is fully automated. No other external knowledge sources are required.
2. Training text can be presented in any order. All of language 1 can be presented, followed by language 2. Alternately, documents from the two languages can be presented in intermixed order.
3. Context vector approach will learn "second order" relationships between the languages used for training. The resulting unified context vector set can be used to identify relationships between words in the two languages.
4. The user can enter multi lingual queries based on tie words as well as non tie words. Because all the text is used during training second order relationships will be formed between non tie words in different languages. As an extreme example if "white" is only used as "white house" and "blanca" is only used as in "casa blanca" the user will be able to query using only "white" and Spanish documents about "casa blanca" will be retrieved. This is not so in the previous approach where the user is limited to using only tie words as query terms.
5. The basic approach described here is extensible and capable of processing more than two languages at once. Additionally, this approach can be utilized for ideographic languages such as Japanese, Chinese and Korean.

The key benefit of the MatchPlus context vector approach is its ability to learn the relationships between words. To simply disregard the relationships contained in the foreign data simply does not make sense. The Symmetric Learning approach exploits the learned relationships without the need to translate the foreign text. The Symmetric Learning approach requires only the translation of a limited number of words (tie words). Furthermore, this operation need only be done once.

The benefits of a multilingual approach to text processing extend well beyond text retrieval. Obviously, text routing and index term assignment could benefit from multilingual technology. Language learning tools could exploit the technology to analyze the relationships between word usage's across languages. Finally, as innovative text visualization techniques are found, multilingual text processing will surely enhance the value of such technology.

5. PRELIMINARY RESULTS OF SYMMETRIC LEARNING TESTS

As stated above, HNC has implemented a limited scope preliminary test of the symmetric learning approach. A preliminary set of 465 tie-words was prepared. This list consisted of words of nearly equivalent meaning in both English and Spanish. Approximately 100 tie-words were selected from the Spanish TREC topics. The balance were selected from high frequency words in the Spanish text. The corpus used for testing consisted of data from three sources as shown in Table 1.

Source	Language	Year	Documents	Size
El Norte	Spanish	1993	395	1.55 MB
TREC AP News	English	1990	69	226 KB
Data Times	English	1993	416	1.37 MB
TOTAL			**880**	**3.145 MB**

Table 1. Bilingual Training Corpus Statistics.

The total number of stems in this test was 32739. The stemmer was disabled for both Spanish and English. The existing MatchPlus learning algorithm was run on the resulting bilingual corpus. When training was complete, a series of stem trees were prepared to assess the nature of the learned relationships. Ideally, one would hope to see both English and Spanish words in the stem trees. The presence of bilingual information in the tree would indicate that the basic approach is viable.

Stem trees were performed for both tie-words and non-tie-words. Based upon the earlier assertion that the presence of bilingual information indicated correct behavior, the true proof of the concept is to demonstrate that bilingual information occurs in stem trees for non-tie-words.

Stem	Doc Freq	Stem Freq	Dot Product
aids	86	578	1.000
aids^patients	12	22	0.699
infected	15	40	0.579
cases	20	35	0.540
hiv	18	49	0.529
tests	26	44	0.528
related	20	26	0.519
humana	17	21	0.517
infectados	14	20	0.500
portadores	16	31	0.500
epidemic	27	56	0.498
virus	67	163	0.493
smndrome	27	33	0.483
inmunodeficiencia	21	26	0.481
infectadas	18	21	0.479
discrimination	22	38	0.463
activists	22	38	0.442
adquirida	26	28	0.442
panel	12	22	0.415
disease	21	53	0.403
sick	31	58	0.368
education	36	49	0.362

Table 2. Stem Tree for Tie Word “Aids”.

Stem	Doc Freq	Word Freq	Dot Product
immunodeficiency	6	6	1.000
hiv	18	49	0.821
virus	67	163	0.730
infected	15	40	0.661
causes	25	35	0.599
humana	17	21	0.557
aids	86	578	0.550
inmunodeficiencia	21	26	0.514
touted	5	5	0.488
smndrome	27	33	0.452
portadores	16	31	0.444
human	107	198	0.433
infectadas	18	21	0.418
positive	20	27	0.417
cases	20	35	0.371
inmuno	6	8	0.366
infectados	14	20	0.360
portadoras	5	5	0.358
deficiencia	8	10	0.351
vih	10	18	0.343
adquirida	26	28	0.340
infection	12	22	0.337
aids^patients	12	22	0.328
test	16	21	0.321
abusers^drug	5	6	0.319

Table 3. Stem Tree for Non-Tie-Word “Immunodeficiency”

Stem	Doc Freq	Word Freq	Dot Product
inmunodeficiencia	21	26	1.000
smndrome	27	33	0.939
adquirida	26	28	0.886
humana	17	21	0.855
inmuno	6	8	0.739
virus	67	163	0.711
deficiencia	8	10	0.607
portadores	16	31	0.528
immunodeficiency	6	6	0.514
infectadas	18	21	0.504
hiv	18	49	0.496
vih	10	18	0.489
aids	86	578	0.481
infectados	14	20	0.390
portadoras	5	5	0.362
causes	25	35	0.332
aids^patients	12	22	0.329
infected	15	40	0.327
causante	5	7	0.314
muerto	15	18	0.309
estima	19	19	0.301
epidemic	27	56	0.298
portador	8	10	0.287
sick	31	58	0.287
provoca	11	12	0.279

Table 4. Stem Tree for Non-Tie-Word "Inmunodeficiencia"

Table 2 shows the stem tree for the tie-word "AIDS". In Spanish, AIDS has the acronym "SIDA" which stands for sindrome inmuno deficiencia adquirida. Also note that HIV in Spanish is VIH for virus inmunodeficiencia humana. Inspection of Table 2 shows that all stems present make sense and that the stem tree captures the contextual similarity across linguistic boundaries. Specifically, the list contains "infected" and "infectadas" and "infectados". Also, the list contains related terms like "portadores' (carriers). These relationships, though not unexpected, bodes well for the potential of the approach. Clearly, for tie-words, the technique will work. The true test is the stem trees for non-tie-words.

Table 3 and Table 4 show the stem trees for the non-tie-words "immuniodeficiency" and "inmunodeficiencia". Clearly, their context of usage in the two languages should be similar and consequently, their stem trees should be similar. Inspection of Table 3 and Table 4 show exactly the type of behavior desired. Indeed, all the correct Spanish terms are present in the stem tree for the English root "immuniodeficiency" and likewise for the Spanish root "inmunodeficiencia". This data suggests that the proposed approach has a very high probability of correctly representing the terms in both Spanish and English in a unified meaning space.

To add one more example of the ability of the context vector approach to identify second order relationships, consider the stem tree for the term "monopoly" shown in Table 5. Notice that MatchPlus correctly detects the relationship between "monopoly" and "pemex" the Mexican national oil company.

Stem	Doc Freq.	Word Freq.	Dot Product
monopoly	20	43	1.000
pemex	44	314	0.639
oil	36	48	0.565
stolen	16	16	0.498
refinacisn	12	23	0.475
reestructuracisn	15	24	0.445
basica	7	9	0.423
petrolera	12	44	0.372
competencia	28	53	0.351
shell	9	18	0.347
reforma	24	46	0.323
privatization	12	25	0.306
exploracisn	9	18	0.305
petrsleos	22	26	0.304
oil^shell	3	5	0.304
petroqummica	12	33	0.302

Table 5. Stem Tree for Tie-Word "Monopoly"

5.1. Multilingual Approach Summary

The success demonstrated on bilingual text characterization strongly suggests that the context vector approach will provide an effective means of providing multilingual information retrieval. Because the preliminary results are so positive, HNC proposes to extend the preliminary demonstration system to Spanish in two steps. The first step for Spanish will be to develop all the required support files and software for the Spanish language including stop lists, stemmers, etc. This will result in a full capability MatchPlus system for Spanish. The second step will be to continue development of an English-Spanish MatchPlus system by augmenting the current tie-word list and to perform a series of engineering experiments. These experiments will identify the sensitivity of system performance to the characteristics of the tie-words chosen.

REFERENCES

[1] Salton, G. (ed.), "The SMART Retrieval System - Experiments in Automatic Document Processing", Prentice-Hall, 1971.

[2] Salton, G., "Another Look at Automatic Text Retrieval Systems", Communications of the ACM, Vol. 20, 1986, pp. 648 - 656.

[3] Salton, G., "Automatic Text Processing", Addison-Wesley, 1989.

[4] Sutcliffe, R., "Distributed Representations in a Text Based Information Retrieval System: A New Way of Using the Vector Space Model", Communications of the ACM, Jan. 1991, pp. 123 - 132.

[5] Koll, M.B., "WEIRD: An Approach to Concept-Based Information Retrieval", SIGIR Forum, Vol. 13, No. 4, Spring 1979, pp. 32 - 50.

[6] Watson, G.S., "Statistics on Spheres", John Wiley and Sons, 1983.

[7] Gallant, S.I., W. R. Caid, et al, "Feedback and Mixing Experiments with MatchPlus", Proceedings TREC-2 Conference, D. Harman, Ed, Gaithersburg, MD. Aug. 1993.

[8] Sasseen, R. V., J. L. Carleton, W. R. Caid, "CONVECTIS: A Context Vector-Based On-Line Indexing System", in *Proceedings IEEE Dual-Use Conference*, 1995.